Issues that could be improved

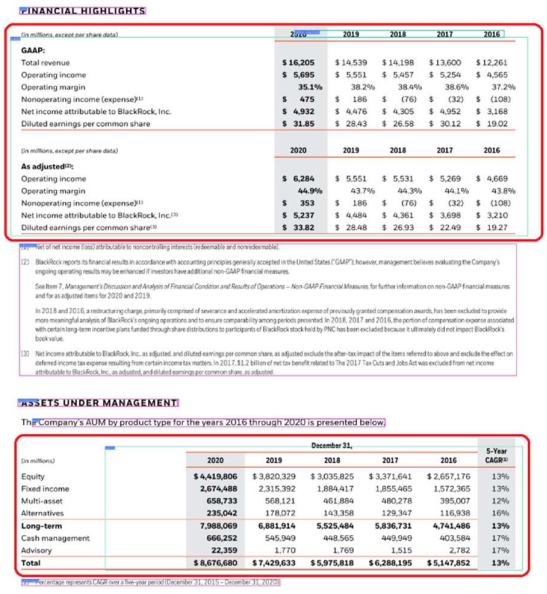
1. table detection model

Layoutparser is not the most accurate model for table detection, if we could find a more powerful model, it would substantially improve the performance of our application. I noticed that [PP-StructureV2] claims in its readme that it is more accurate and faster than layoutparser.
(https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.6/ppstructure/docs/PP-StructureV2_introduction.md)

On the PubLayNet dataset, the performance comparison with other methods is shown in the following table. It can be seen that compared to the layoutparser, a layout analysis tool based on Detectron2, our model is about 5% more accurate and predicts about 69 times faster.

model	mAP	CPU prediction takes time
layoutparser (Detectron2)	88.98%	2.90s
PP-StructureV2 (PP-PicoDet)	**94.00%**	41.20ms

So, I tried to use PP-Structure to detect the tables in pages 8 to 18 of "blackrock-2020-annual-report.pdf". The result shows that it can detect more tables than layoutparser, but it will also mistake some paragraphs for tables.
In this screenshot of page 16, the table boxed by the red rectangle were detected by both methods, blue one was only detected by PP-Structure, and purple is where PP-Structure mistake text for tables.
If we look closely, we can see that there are also some light blue rectangles, they are the areas that PP-structures consider as tables. We can find that most of them are slightly smaller than the real table area, and that will cause extraction problem.



FINANCIAL HIGHLIGHTS

(in millions, except per share data)	2020	2019	2018	2017	2016
GAAP:					
Total revenue	**$ 16,205**	$ 14,539	$ 14,198	$ 13,600	$ 12,261
Operating income	**$ 5,695**	$ 5,551	$ 5,457	$ 5,254	$ 4,565
Operating margin	**35.1%**	38.2%	38.4%	38.6%	37.2%
Nonoperating income (expense)[1]	**$ 475**	$ 186	$ (76)	$ (32)	$ (108)
Net income attributable to BlackRock, Inc.	**$ 4,932**	$ 4,476	$ 4,305	$ 4,952	$ 3,168
Diluted earnings per common share	**$ 31.85**	$ 28.43	$ 26.58	$ 30.12	$ 19.02
(in millions, except per share data)	2020	2019	2018	2017	2016
As adjusted[2]:					
Operating income	**$ 6,284**	$ 5,551	$ 5,531	$ 5,269	$ 4,669
Operating margin	**44.9%**	43.7%	44.3%	44.1%	43.8%
Nonoperating income (expense)[1]	**$ 353**	$ 186	$ (76)	$ (32)	$ (108)
Net income attributable to BlackRock, Inc.[3]	**$ 5,237**	$ 4,484	$ 4,361	$ 3,698	$ 3,210
Diluted earnings per common share[3]	**$ 33.82**	$ 28.48	$ 26.93	$ 22.49	$ 19.27

Net of net income (loss) attributable to noncontrolling interests redeemable and nonredeemable.

(2) BlackRock reports its financial results in accordance with accounting principles generally accepted in the United States ("GAAP"); however, management believes evaluating the Company's ongoing operating results may be enhanced if investors have additional non-GAAP financial measures.

See Item 7, *Management's Discussion and Analysis of Financial Condition and Results of Operations – Non-GAAP Financial Measures*, for further information on non-GAAP financial measures and for as adjusted items for 2020 and 2019.

In 2018 and 2016, a restructuring charge, primarily comprised of severance and accelerated amortization expense of previously granted compensation awards, has been excluded to provide more meaningful analysis of BlackRock's ongoing operations and to ensure comparability among periods presented. In 2018, 2017 and 2016, the portion of compensation expense associated with certain long-term incentive plans funded through share distributions to participants of BlackRock stock held by PNC has been excluded because it ultimately did not impact BlackRock's book value.

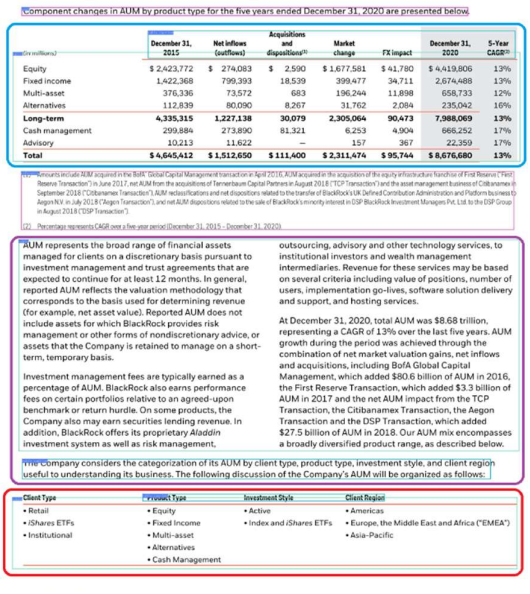
(3) Net income attributable to BlackRock, Inc., as adjusted, and diluted earnings per common share, as adjusted exclude the after-tax impact of the items referred to above and exclude the effect on deferred income tax expense resulting from certain income tax matters. In 2017, $1.2 billion of net tax benefit related to The 2017 Tax Cuts and Jobs Act was excluded from net income attributable to BlackRock, Inc., as adjusted, and diluted earnings per common share, as adjusted.

ASSETS UNDER MANAGEMENT

The Company's AUM by product type for the years 2016 through 2020 is presented below:

(in millions)	December 31, 2020	2019	2018	2017	2016	5-Year CAGR[2]
Equity	**$ 4,419,806**	$ 3,820,329	$ 3,035,825	$ 3,371,641	$ 2,657,176	13%
Fixed income	**2,674,488**	2,315,392	1,889,417	1,855,465	1,572,365	13%
Multi-asset	**658,733**	568,121	461,884	480,278	395,007	12%
Alternatives	**235,042**	178,072	143,358	129,347	116,938	16%
Long-term	**7,988,069**	**6,881,914**	**5,525,484**	**5,836,731**	**4,741,486**	**13%**
Cash management	**666,252**	545,949	448,565	449,949	403,584	17%
Advisory	**22,359**	1,770	1,769	1,515	2,782	17%
Total	**$ 8,676,680**	**$ 7,429,633**	**$ 5,975,818**	**$ 6,288,195**	**$ 5,147,852**	**13%**

(2) Percentage represents CAGR over a five-year period (December 31, 2015 – December 31, 2020).



Component changes in AUM by product type for the five years ended December 31, 2020 are presented below:

(in millions)	December 31, 2015	Net inflows (outflows)	Acquisitions and dispositions[1]	Market change	FX impact	December 31, 2020	5-Year CAGR[2]
Equity	$ 2,423,772	$ 274,083	$ 2,590	$ 1,677,581	$ 41,780	$ 4,419,806	13%
Fixed income	1,422,368	799,393	18,539	399,477	34,711	2,674,488	13%
Multi-asset	376,336	73,572	683	196,244	11,898	658,733	12%
Alternatives	112,839	80,090	8,267	31,762	2,084	235,042	16%
Long-term	**4,335,315**	**1,227,138**	**30,079**	**2,305,064**	**90,473**	**7,988,069**	**13%**
Cash management	299,884	273,890	81,321	6,253	4,904	666,252	17%
Advisory	10,213	11,622	–	157	367	22,359	17%
Total	**$ 4,645,412**	**$ 1,512,650**	**$ 111,400**	**$ 2,311,474**	**$ 95,744**	**$ 8,676,680**	**13%**

(1) Amounts include AUM acquired in the BofA Global Capital Management transaction in April 2010, AUM acquired in the acquisition of the equity infrastructure franchise of First Reserve ("First Reserve Transaction") in June 2017, net AUM from the acquisitions of Tennenbaum Capital Partners in August 2018 ("TCP Transaction") and the asset management business of Citibanamex in September 2018 ("Citibanamex Transaction"), AUM reclassifications and net dispositions related to the transfer of BlackRock's UK Defined Contribution Administration and Platform business to Aegon N.V. in July 2018 ("Aegon Transaction"), and net AUM dispositions related to the sale of BlackRock's minority interest in DSP BlackRock Investment Managers Pvt. Ltd. to the DSP Group in August 2018 ("DSP Transaction").

(2) Percentage represents CAGR over a five-year period (December 31, 2015 – December 31, 2020).

AUM represents the broad range of financial assets managed for clients on a discretionary basis pursuant to investment management and trust agreements that are expected to continue for at least 12 months. In general, reported AUM reflects the valuation methodology that corresponds to the basis used for determining revenue (for example, net asset value). Reported AUM does not include assets for which BlackRock provides risk management or other forms of nondiscretionary advice, or assets that the Company is retained to manage on a short-term, temporary basis.

Investment management fees are typically earned as a percentage of AUM. BlackRock also earns performance fees on certain portfolios relative to an agreed-upon benchmark or return hurdle. On some products, the Company also may earn securities lending revenue. In addition, BlackRock offers its proprietary Aladdin investment system as well as risk management, outsourcing, advisory and other technology services, to institutional investors and wealth management intermediaries. Revenue for these services may be based on several criteria including value of positions, number of users, implementation go-lives, software solution delivery and support, and hosting services.

At December 31, 2020, total AUM was $8.68 trillion, representing a CAGR of 13% over the last five years. AUM growth during the period was achieved through the combination of net market valuation gains, net inflows and acquisitions, including BofA Global Capital Management, which added $80.6 billion of AUM in 2016, the First Reserve Transaction, which added $3.3 billion of AUM in 2017 and the net AUM impact from the TCP Transaction, the Citibanamex Transaction, the Aegon Transaction and the DSP Transaction, which added $27.5 billion of AUM in 2018. Our AUM mix encompasses a broadly diversified product range, as described below.

The Company considers the categorization of its AUM by client type, product type, investment style, and client region useful to understanding its business. The following discussion of the Company's AUM will be organized as follows:

Client Type	Product Type	Investment Style	Client Region
• Retail	• Equity	• Active	• Americas
• iShares ETFs	• Fixed Income	• Index and iShares ETFs	• Europe, the Middle East and Africa ("EMEA")
• Institutional	• Multi-asset		• Asia-Pacific
	• Alternatives		
	• Cash Management		

Therefore, the choice depends on our needs. If we want to find as many table as possible, we should choose PP-Structure. Otherwise, the layoutparser is a better choice if we want the extracted tables have a higher accuracy rate.

2. draw lines between rows and columns in borderless tables

PP-structure sometimes has trouble extracting borderless tables, so I was wondering if drawing split lines for the table in advance and then extracting the pre-processed table image could improve the extraction results.
Unfortunately, I didn't find many ways to add borders to the table. cv2 have function like findContours(), However, it is not as accurate as the SLANet model used by PP-structure.
As we can see, this function cannot recognize the original border or cells that occupy many rows, it is just adding split lines between every two rows of objects.

	Shares Issued				Shares Outstanding		
	Common Shares	Treasury Common Shares	Series B Preferred	Series C Preferred	Common Shares	Series B Preferred	Series C Preferred
December 31, 2017	171,252,185	(11,275,070)	823,188	246,522	159,977,115	823,188	246,522
Shares repurchased	—	(3,511,603)	—	—	(3,511,603)	—	—
Net issuance of common shares related to employee stock transactions	—	1,087,989	—	—	1,087,989	—	—
PNC LTIP capital contribution	—	—	—	(103,064)	—	—	(103,064)
December 31, 2018	171,252,185	(13,698,684)	823,188	143,458	157,553,501	823,188	143,458
Shares repurchased	—	(4,018,905)	—	—	(4,018,905)	—	—
Net issuance of common shares related to employee stock transactions	—	841,184	—	—	841,184	—	—
PNC LTIP capital contribution	—	—	—	(143,458)	—	—	(143,458)
December 31, 2019	171,252,185	(16,876,405)	823,188	—	154,375,780	823,188	—
Shares repurchased	—	(3,445,554)	—	—	(3,445,554)	—	—
Net issuance of common shares related to employee stock transactions	—	779,471	—	—	779,471	—	—
Exchange of preferred shares series B for common shares	823,188	—	(823,188)	—	823,188	(823,188)	—
December 31, 2020	**172,075,373**	**(19,542,488)**	**—**	**—**	**152,532,885**	**—**	**—**

Even worse, sometimes it fails to detect rows and consider the entire table as one single row.

(in millions)	2020	2019	2018
Beginning balance	**$(571)**	$(691)	$(432)
Foreign currency translation adjustments[1]	**234**	120	(253)
Reclassification as a result of adoption of accounting guidance	**—**	—	(6)
Ending balance	**$(337)**	$(571)	$(691)

In fact, PP-structure's recognition result of cells is a combination of the coordinates, recognition result of a single row, and the coordinates of the cells, and it may be hard to find a better open source “line-drawer”, if even PP-structure cannot recognize the cells well.
But if we can find a more accurate table recognizer, it will greatly improve the performance of our program.